Chapter 11: Two Variable Regression Analysis

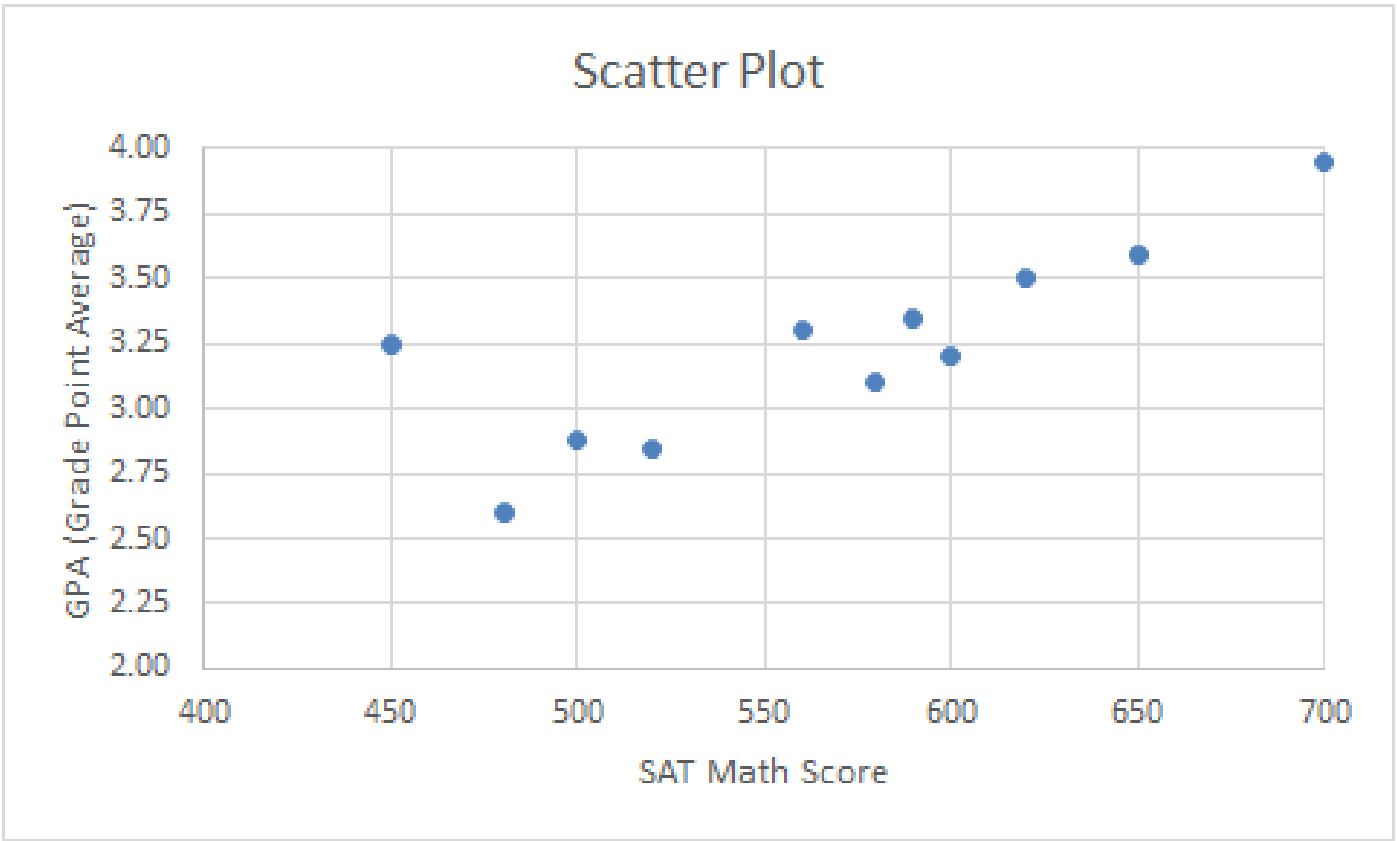
Department of Mathematics
Izmir University of Economics

Week 14-15
2014-2015

In this chapter, we will focus on

- linear models and extend our analysis to relationships between variables,
- the definitions of *SSR*, *SSE*, *SST*, and coefficient of determination,
- ANOVA tables, and
- hypothesis test for correlation between two variables.

In Chapter 1, we learned how the relationship between two variables can be described by using scatter plots to see the picture of the relationship.



Moreover, in Chapter 2, we learned that the covariances and correlation coefficients provide numerical measures of that relationship.

- A population covariance is

$$Cov(X, Y) = \sigma_{X,Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- A sample covariance is

$$Cov(X, Y) = s_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- A population correlation coefficient is

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

- A sample correlation coefficient is

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

Remark: ρ and r are always between -1 and 1.

Here we can approximate the relationship by a linear equation

$$Y = \beta_0 + \beta_1 X,$$

where

Y is the dependent variable: the variable we wish to explain (also called the endogenous variable)

X is the independent variable: the variable used to explain the dependent variable (also called the exogenous variable)

β_0 is the intercept: where the line cuts Y-axis.

β_1 is the slope of the line. (This slope is very important because it indicates the change in Y-variable when the variable X changes.)

In order to find the best linear relationship between Y and X, we use Least Square Regression Technique. This technique computes estimates for β_0 and β_1 as b_0 and b_1.

The Least Squares Regression line based on sample data is;

$$\hat{y} = b_0 + b_1 x,$$

where $\mathbf{b_1}$ is the **slope** of the line given by

$$b_1 = \frac{s_{X,Y}}{s_x^2} = r\frac{s_y}{s_x}$$

and $\mathbf{b_0}$ is the y-**intercept**

$$b_0 = \bar{y} - b_1\bar{x}.$$

Here $\hat{y}$ is called as the **estimated value**.

Example: An instructor in a statistics course set a final examination and also required the students to do a data analysis project. For a random sample of 10 students, the scores obtained are shown in the table. Find the sample correlation between the examination and project scores. Estimate a linear regression of project scores on exam scores.

Examination	81	62	74	78	93	69	72	83	90	84
Project	76	71	69	76	87	62	80	75	92	79

Example: Complete the following for the (x, y) pairs of data points

$$(1, 5), (3, 7), (4, 6), (5, 8), (7, 9).$$

a) Compute b_1.

b) Compute b_0.

c) What is the equation of the regression line?

Linear regression model

Linear regression population equation model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where β_0 and β_1 are the population model coefficients and ε is a random error term.

Standard Assumptions:

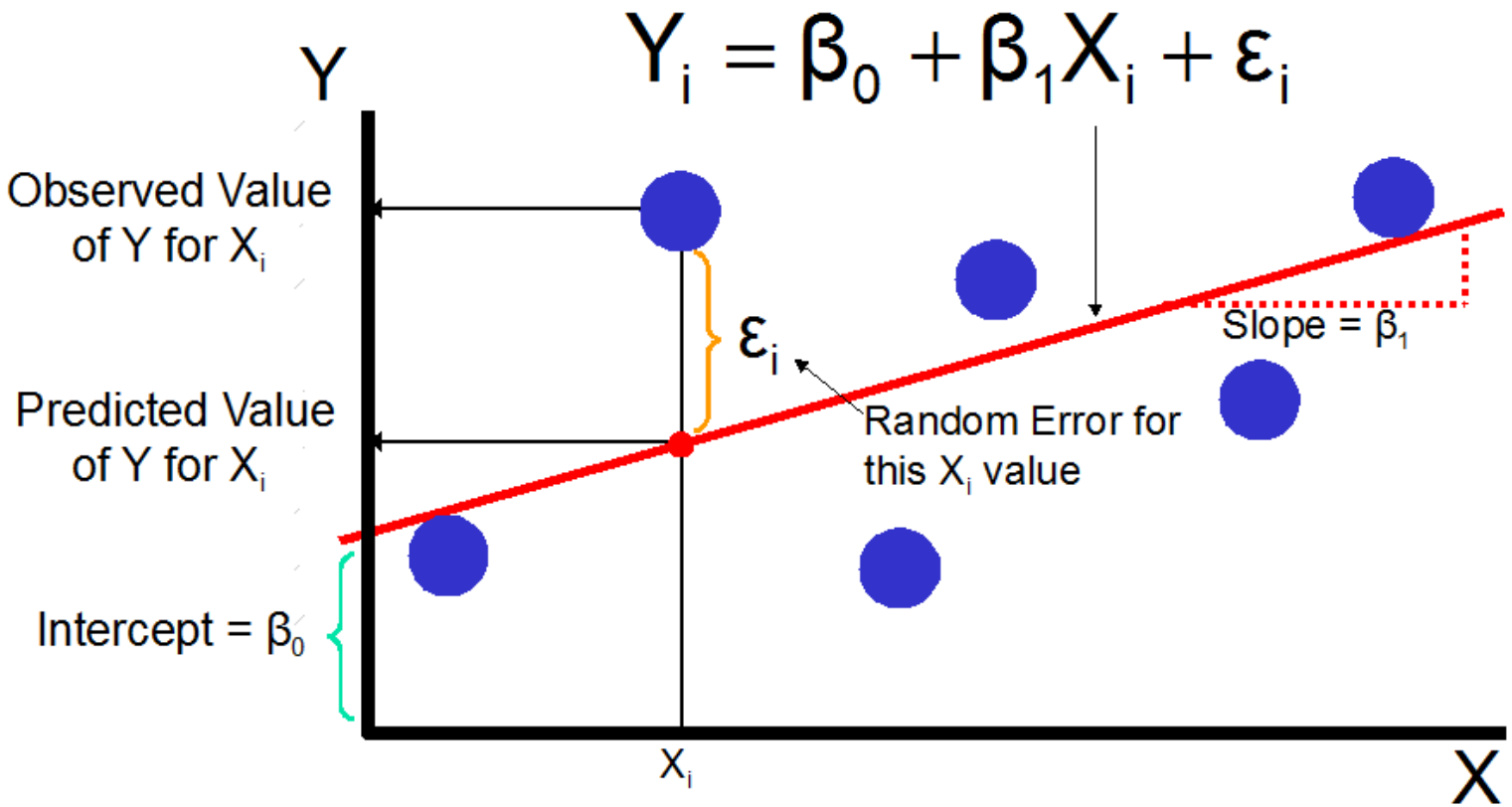
- The true relationship form is linear (Y is a linear function of X plus some random error).
- The error terms, ε_i are independent of the x values.
- The error terms are random variables with mean 0 and constant variance, σ^2:

$$E(\varepsilon_i) = 0 \qquad E(\varepsilon_i^2) = \sigma^2.$$

- The random error terms, ε_i, are not correlated with one another.

$$E(\varepsilon_i \varepsilon_j) = 0 \qquad \text{for all} \qquad i \neq j.$$

Explaining Coefficients



Least squares coefficient estimators

Estimates:

- $\widehat{\sigma}_{x,y} = s_{x,y}$
- $\widehat{\rho} = r$
- $\widehat{\beta}_0 = b_0$
- $\widehat{\beta}_1 = b_1$
- $\widehat{\varepsilon}_i = e_i$

Estimated model (based on a random sample):

$$y_i = b_0 + b_1 x_i + e_i$$

and we call the *error* as residual:

$$e_i = y_i - \widehat{y}_i = y_i - (b_0 + b_1 x_i)$$

Least squares

- The coefficients b_0 and b_1 are found so that

$$SSE = \sum e_i^2$$

  is minimized.
- By using some calculus, we get

$$b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

  Alternatively, we use

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

Example: For a sample of 20 monthly observations, a financial analyst wants to regress the percentage rate of return (Y) of the common stock of a corporation on the percentage rate of return (X) of the Standard and Poor's 500 Index. The following information is available:

$$\sum_{i=1}^{20} y_i = 22.6 \quad \sum_{i=1}^{20} x_i = 25.4 \quad \sum_{i=1}^{20} x_i^2 = 145.7 \quad \sum_{i=1}^{20} x_i y_i = 150.5$$

a) Estimate the linear regression of Y on X.

b) Interpret the slope of the sample regression line.

c) Interpret the intercept of the sample regression line.

Analysis of Variance (ANOVA)

Note that the *total variance* is computed via $\sum(y_i - \bar{y})^2$.

This can be divided into two parts which will lead us further analysis of regression:

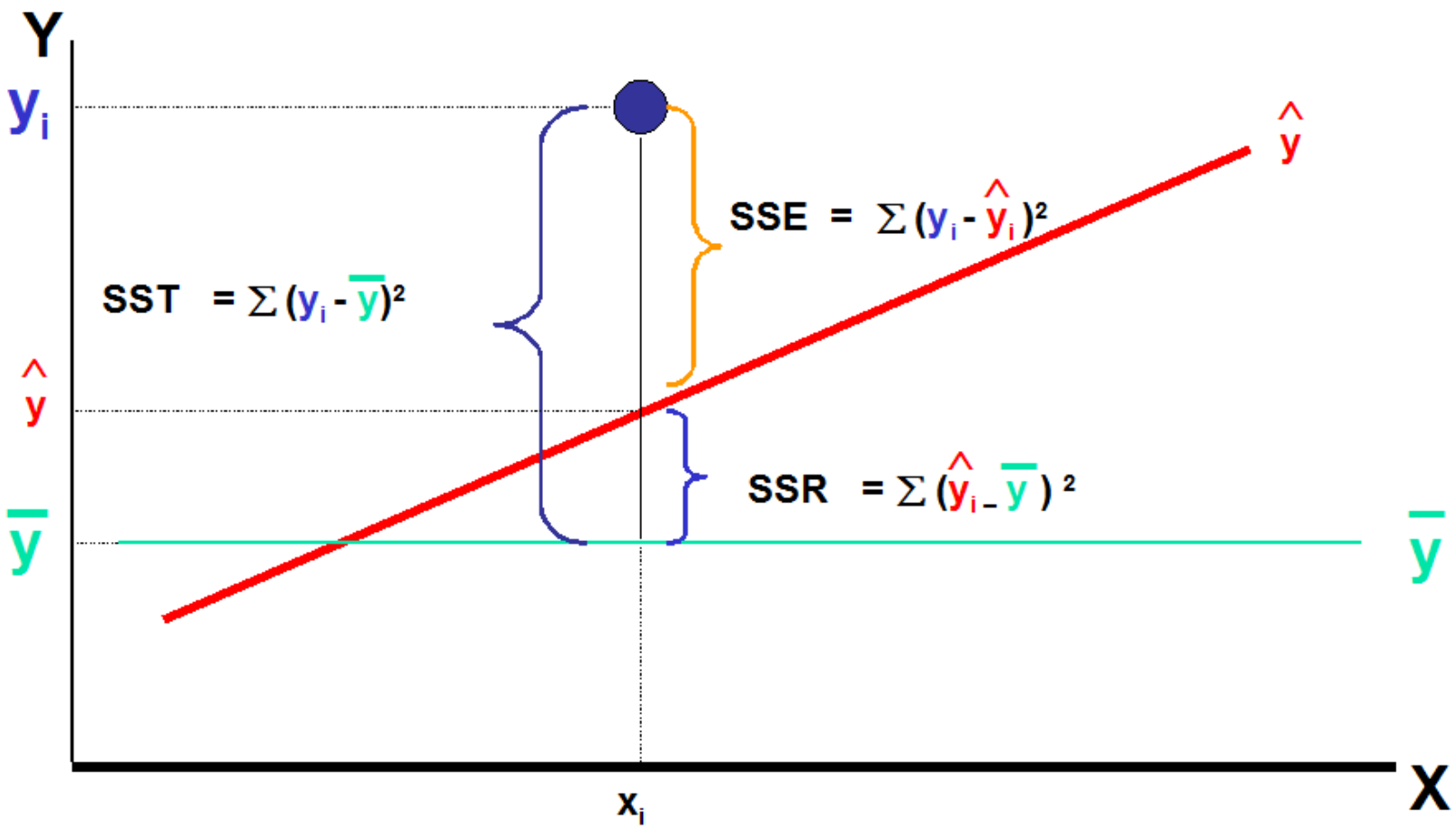
$$\sum(y_i - \bar{y})^2 = \sum(y_i - \widehat{y}_i)^2 + \sum(\widehat{y}_i - \bar{y})^2$$

Here we call them

- Total sum of squares $SST = \sum(y_i - \bar{y})^2$,
- Regression sum of squares $SSR = \sum(\widehat{y}_i - \bar{y})^2 = b_1^2 \sum(x_i - \bar{x})^2$, and
- Error sum of squares $SSE = \sum(y_i - \widehat{y}_i)^2 = \sum e_i^2$.

So that, $SST = SSR + SSE$.

Explaining squares



Coefficient of determination

Note that for given sample, we cannot generally control SST but we may control SSR and SSE when defining b_0 and b_1 where we tried to minimize SSE. So, it might be a good guess to look for the ratio SSR/SST which may represent the success of regression:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Note that we always have $0 \leq R^2 \leq 1$.
- R^2 can be used to compare two regression models.
- Important: $R^2 = r^2$.

Model Error Variance is given by $\widehat{\sigma}^2 = s_e^2 = \frac{SSE}{n-2}$, where s_e is called standard error of the regression.

Example: Compute SSR, SSE, s_e^2 and the coefficient of determination given the following statistics computed from a random sample of pairs of X and Y observations:

$$\sum (y_i - \bar{y})^2 = 100000; \quad r^2 = 0.50; \quad n = 52.$$

Statistical inference: Hypothesis tests and confidence intervals

If the standard least squares assumptions hold, then b_1 is an unbiased estimator for β_1, that is,

$$\widehat{\beta}_1 = b_1.$$

Moreover, its population variance is

$$\sigma^2_{b_1} = \frac{\sigma^2}{(n-1)s_x^2}$$

and its unbiased sample variance estimator is

$$s^2_{b_1} = \frac{s_e^2}{(n-1)s_x^2}.$$

There is a similar but more complicated formula for b_0 which we will not give details.

Hypothesis test for slope

In order to test hypothesis about β_1 and give more detailed estimation such as confidence interval, we need to define the corresponding statistics and distributions.

- Two-tailed hypothesis test:

  $$\text{Test } H_0 : \beta_1 = \beta_1^* \text{ against } H_1 : \beta_1 \neq \beta_1^*$$

  with test statistic

  $$t = \frac{b_1 - \beta_1^*}{s_{b_1}},$$

  which follows a Student's t-distribution with $(\mathbf{n} - \mathbf{2})$ degrees of freedom and decision rule

  $$\text{Reject } H_0 \text{ if } t \leq -t_{n-2,\frac{\alpha}{2}} \text{ or } t \geq t_{n-2,\frac{\alpha}{2}}.$$

- One-tailed versions are tested analogously:

  $$\text{For } H_1 : \beta_1 > \beta_1^*, \text{ we reject } H_0 \text{ if } t \geq t_{n-2,\alpha},$$

  $$\text{For } H_1 : \beta_1 < \beta_1^*, \text{ we reject } H_0 \text{ if } t \leq -t_{n-2,\alpha}.$$

Confidence interval

Similarly, we can describe the slope (β_1) by giving a confidence interval which will also reflect the significance level:

$$CI: \quad b_1 - t_{n-2,\frac{\alpha}{2}} s_{b_1} < \beta_1 < b_1 + t_{n-2,\frac{\alpha}{2}} s_{b_1}$$

Example: Given the simple regression model

$$Y = \beta_0 + \beta_1 X$$

and the regression results that follow, test the null hypothesis that the slope coefficient is 0 versus the alternative hypothesis of greater than zero using probability of Type I error equal to 0.05, that is, $\alpha = 0.05$. Also find the 95% confidence interval for the slope coefficient.

a) A random sample of size $n = 38$ with $b_1 = 5$ and $s_{b_1} = 2.1$.

b) A random sample of size $n = 29$ with $b_1 = 6.7$ and $s_{b_1} = 1.8$.

F distribution and F-test

For independent and normally distributed populations, we define a new random variable

$$F = \frac{\frac{s_x^2}{\sigma_x^2}}{\frac{s_y^2}{\sigma_y^2}},$$

where s_x^2 and s_y^2 are sample variances.

This random variable has an F distribution with $(n_x - 1)$ numerator degrees of freedom and $(n_y - 1)$ denominator degrees of freedom. (In short, we write F_{v_1, v_2}.)

In order to find critical values (cutoff points), we need to define a test statistic:

$$F = \frac{s_x^2}{s_y^2},$$

where we take $s_x > s_y$ (and hence $F > 1$).

- We can use F test to conclude **two-sided hypothesis tests**. We can test the hypothesis

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0$$

with test statistic

$$F = \frac{MSR}{MSE},$$

where $MSR = \dfrac{SSR}{k}$ is called the mean square for regression (Note that, k is the number of independent variables. So, for simple regression $k = 1$.) and $MSE = \dfrac{SSE}{n-2}$ is called the mean square for error. That is, in short we have $F = \dfrac{SSR}{s_e^2}$.

The decision rule is

$$\text{Reject } H_0 \text{ if } F \geq F_{1,n-2,\alpha}.$$

Note: Although F test requires two-sided hypothesis test, we use α not $\frac{\alpha}{2}$!!!

Example: Test at 5% significance level against two-sided alternative the null hypothesis that the slope of the population regression line is 0, where $SST = 128000$, $n = 25$, and $r = 0.69$.

Hypothesis test for correlation

In order to test that there are no linear relations, we test $H_0 : \rho = 0$ with test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which follows a Student's t-distribution with $\mathbf{n-2}$ degrees of freedom.

- The decision rule for one-sided alternative $H_1 : \rho > 0$ is

  Reject H_0 if $t > t_{n-2,\alpha}$.

- The decision rule for one-sided alternative $H_1 : \rho < 0$ is

  Reject H_0 if $t < -t_{n-2,\alpha}$.

- The decision rule for two-sided alternative $H_1 : \rho \neq 0$ is

  Reject H_0 if $t < -t_{n-2,\frac{\alpha}{2}}$ or $t > t_{n-2,\frac{\alpha}{2}}$.

Example: Test the null hypothesis

$$H_0 : \rho = 0,$$

versus

$$H_1 : \rho \neq 0,$$

given the following: A sample correlation of 0.60 for a random sample of size $n = 25$.

Example: For a random sample of 353 high school teachers the sample correlation between annual raises and teaching evaluations was found to be 0.11. Test the null hypothesis that these quantities are uncorrelated in the population against the alternative that the population correlation is positive.

Example: Doctors are interested in the relationship between the dosage of a medicine and the time required for a patient's recovery. The following table shows, for a sample of five patients, dosage levels and recovery times. These patients have similar characteristics except for medicine dosages.

Dosage level	1.2	1.0	1.5	1.2	1.4
Recovery time	25	40	10	27	16

a) Estimate the linear regression of recovery time on dosage level.

b) Find and interpret a 90% confidence interval for the slope of the population regression line.

c) Would the sample regression derived in part **a)** be useful in predicting recovery time for a patient given 2.5 grams of this drug?

Example: For a random sample of 526 firms, the sample correlation between the proportion of a firm's officers who are directors and a risk-adjusted measure of return on the firm's stock was found to be 0.1398. Test against a two-sided alternative the null hypothesis that the population correlation is 0.

Example: Based on a sample of 30 observations, the population regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

was estimated. The least squares estimates obtained were

$$b_0 = 10.1 \qquad b_1 = 8.4$$

The regression and error sums of squares were

$$SSR = 128 \qquad SSE = 286$$

a) Find and interpret the coefficient of determination.

b) Test at the 1% significance level against a two-sided alternative the null hypothesis that β_1 is 0.

Example: An analyst believes that the only important determinant of banks' returns on assets (Y) is the ratio of loans to deposits (x). For a random sample of 20 banks the sample regression line

$$Y = 0.97 + 0.47x$$

was obtained with coefficient of determination 0.720.

a) Find the sample correlation between returns on assets and the ratio of loans to deposits.

b) Test against a two-sided alternative at the 5% significance level the null hypothesis of no linear association between the returns and the ratio.

ANOVA from MS Excel

	A	B	C	D	E	F	G	H	I	J
1	**Exam**	**Project**		SUMMARY OUTPUT						
2	81	76								
3	62	71		*Regression Statistics*						
4	74	69		Multiple R	0.775857					
5	78	76		R Square	0.601955					
6	93	87		Adjusted R Square	0.552199					
7	69	62		Standard Error	5.765566					
8	72	80		Observations	10					
9	83	75								
10	90	92		ANOVA						
11	84	79			*df*	*SS*	*MS*	*F*	*Significance F*	
12				Regression	1	402.16599	402.16599	12.09822	0.00834	
13				Residual	8	265.93401	33.24175			
14				Total	9	668.10000				
15										
16					*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17				Intercept	21.80204	15.88817	1.37222	0.20724	-14.83618	58.44026
18				Exam	0.69845	0.20080	3.47825	0.00834	0.23539	1.16150
19										

Example: The commercial division of a real estate firm conducted a study to determine the extent of the relationship between annual gross rents ($1000s) and the selling price ($1000s) for apartment buildings. Data were collected on several properties sold, and Excel's Regression tool was used to develop an estimated regression equation. A portion of the Excel output follow.

a) How many apartment buildings were in the sample?

b) Write the estimated regression equation.

c) Use the *t* test to determine whether the selling price is related to annual gross rents.

d) Use the *F* test to determine whether the selling price is related to annual gross rents.

e) Estimate the selling price of an apartment building with gross annual rents of $50,000.

ANOVA

	df	SS	MS	F
Regression	1	41585.3		
Residual	7			
Total	8	51984.1		

	Coefficients	Standard Error	*t* Stat
Intercept	20.000	3.2213	6.21
Annual Gross Rents	7.210	1.3626	5.29